[19]中华人民共和国专利局

[51]Int.Cl[6]
H04L 12/12
H04L 12/24

[12] 发明专利申请公开说明书

[21] 申请号 97115437.6

[43]公开日 1998 年 3 月 4 日

[11] 公开号 CN 1175147A

[22]申请日 97.7.22

[30]优先权

[32]96.8.23 [33]US[31]701939

[71]申请人 国际商业机器公司

地址 美国纽约州

[72]发明人 C·R·阿萨纳修奥

G·S·戈尔德茨米德特

G·D·H·亨特

S·E·史密斯

[74]专利代理机构 中国专利代理(香港)有限公司

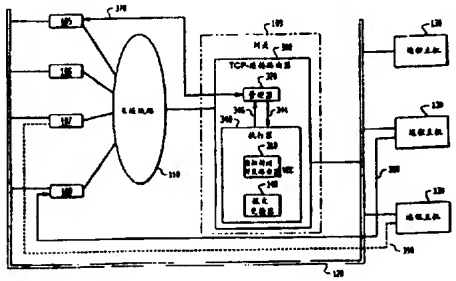
代理人 王 勇 陈景峻

权利要求书 3 页 说明书 13 页 附图页数 12 页

[54]发明名称 可恢复的虚拟封闭群集

[57]摘要

TCP 连接路由器执行封闭群集，其方法是将每个封闭群集划分成若干虚拟封闭群集，根据按照一个可配置策略产生的当前服务器负荷量度，在虚拟封闭群集中动态分配输入连接。在一个实施例中，连接路由器支持群集的动态配置，允许透明地进行故障恢复，为虚拟封闭群集的用户提供不间断服务。



权　利　要　求　书

1．一种为输入报文路由选择通过一个计算机节点群集的边界的方法，该群集与一个或多个网络连接，该方法包括以下步骤：

在端口类型报文的报文首标中寻找并读取端口号和目标地址；

根据该目标地址，选择一个计算机节点子集；

根据该端口号选择一个功能，该功能从该子集中的多个可能目的地中为该报文确定一个路由选择目的地，路由选择目的地是该子集中的一个计算机节点；

发送报文至路由选择目的地；以及

在为报文选择通过边界的路由的同时，动态改变子集中至少一个成员以及子集的数目。

2．权利要求1所述方法，其特征是，进行选择所依据的是报文首标中的端口号和协议标识符。

3．权利要求1所述方法，其特征是，子集的个数由一个监控功能动态改变。

4．权利要求1所述方法，其特征是，子集的成员由该监控功能动态改变。

5．权利要求3所述方法，其特征是，所述改变至少包括替换该子集的成员和增加该子集的成员之一。

6．透明地为输入报文在计算机节点群集上选择路由的网关，其中该群集连接到一个或多个网络，该网关包括：

在端口类型报文的报文首标中查找、读取端口号和目标地址，并根据该目标地址选择一个计算机节点子集的装置；

根据该端口号选择功能的装置，其中该功能为该报文在该子集中多个可能目的地中确定一个路由选择目的地，所述路由选择目的地是子集中的一个计算机节点；

动态改变子集中至少一个成员以及子集的数目的装置。

7．权利要求6所述系统，其特征是，子集的数目是由一个监控功能动态改变的；

8．权利要求7所述系统，其特征是，子集的成员是由该监控功能动态改变的；

9．权利要求7所述系统，其特征是，所述改变至少包括替换该子集的

成员或增加该子集的成员之一。

10．为输入报文在计算机节点群集的边界节点上选择路由的方法，其中该群集连接到一个或多个网络，该方法包括以下步骤：

在该边界节点上，

从端口类型报文的报文首标中查找、读取端口号；

根据该端口号选择一个功能，该功能为该报文在多个可能目的地中确定一个路由选择目的地，所述路由选择目的地是该群集中的一个计算机节点；

检测边界节点的故障；

检测到故障后，将群集中每个节点的各状态信息子集传送到一个替换边界节点；

在替换边界节点处

接收群集中每个节点的状态信息子集；

用该状态信息重建该边界节点发生故障之前的运行状态，使替换边界节点按该边界节点在发生故障之前原本应有的方式对报文进行分配。

11．权利要求10所述方法，其特征是，进一步包括以下步骤：

从端口类型报文的报文首标中查找、读取端口号和目标地址；

根据该目标地址选择一个计算机节点子集；

其中该路由选择目的地只从该子集中选择。

12．恢复计算机节点群集的边界节点故障的系统，包括：

一种装置，它在该边界节点上从端口类型报文的报文首标中查找、读取端口号，根据该端口号选择一个功能，该功能为该报文在多个可能目的地中确定一个路由选择目的地，其中路由选择目的地是该群集中的一个计算机节点；

检测边界节点的故障的装置；

一个替换边界节点，它包括在检测到发生故障后，接收群集中每个节点状态信息子集的装置，以及用该状态信息子集重建该边界节点发生故障之前的运行状态，使替换边界节点按该边界节点在发生故障之前原本应有的方式对报文进行分配的装置。

13．用于计算机节点群集的边界节点，它包括：

一种装置，它从端口类型报文的报文首标中查找、读取端口号，根据该端口号选择一个为该报文在多个可能目的地中确定一个路由选择目的地的

功能，其中路由选择目的地是该群集中的一个计算机节点；

在群集的活动边界节点发生故障后，接收群集中每个节点状态信息子集的装置，以及用该状态信息子集重建该边界节点发生故障之前的运行状态，使替换边界节点按该活动边界节点在发生故障之前原本应有的方式对报文进行分配的装置。

说　明　书

可恢复的虚拟封闭群集

本发明涉及网络计算领域。更确切地说，本发明涉及支持一组远程服务的计算机群集。

封闭群集(Encapsulated Cluster)（ EC ）的特征是一个连接路由器（ Connection-Router ）节点和提供一组服务（例如 Web 服务、 NFS 、等等）的多个服务器主机。美国专利 5,371,852《使一个计算机群集看起来像计算机网络上的一个主机的方法和设备》描述了一例提供封闭群集的系统。

远程用户用基于例如 TCP/IP （如 HTTP ）的协议向封闭群集提出服务请求。为每个请求的服务时间不等，这取决于服务的类型以及相应的服务器应用程序（ applications ）的可获得性。所以，如果草率地对连接进行分配，很快会产生分配偏斜，降低了对可用的封闭群集资源的利用率，导致对用户请求不必要的延迟。

现有技术水平表明，在定标服务器（ scaling servers ）上，存在许多性能问题。可参阅的文章有—例如—《 NCSA 的环球网络服务器: 设计与性能》（ NCSA's World Wide Web Server : Design and Performance ），该文发表在 1995 年 11 月出版的第 28 卷 11 期 IEEE 计算机杂志（ 68 ~ 74 页）上。以用循环式 DNS 支持 Web 服务器（即 httpd 守护程序）的封闭群集为例。服务器通过 http 提供对视频信号流、数据库检索以及静态 web 页面的访问服务。根据所提供的服务的类型以及所涉及的实际内容，服务器为每个请求服务的时间有很大差别。例如，一次复杂的数据库检索的时间可能会比提供一个预先装载的静态 HTML 页面的时间多几个数量级。服务请求处理时间的这种不平衡，经常引起对服务器群集利用上的偏斜。循环式 DNS 的有关问题，在 Kwa 等人所著的《用户对 NCSA 的环球网络服务器的访问模式》（ User Access Patterns to NCSA's Worldwide Web Server ）一文中有所论述，该文登载于美国伊利诺伊斯大学 Urbana — Champaign 分校计算机科学系 1995 年 2 月的“技术报告 UIUCDSD — R — 95 — 1394 ”中。

现有技术水平表明，需要有进行动态资源分配的方法。可参阅的文章有—例如—《通过代表评估管理决策》（ Evaluating Management Decisions via Delegation ），该文作者是 German Goldszmidt 和 Yechian Yemeni，发表于 1993 年 4 月在美国加州旧金山召开的“第三届集成网络管理国际研讨会”。

封闭群集一般来说是在多个主机上以一个整体系统形象提供一系列服务。然而，系统的实际配置情况可能会要求分配服务要遵守特定的、可能是动态的用户原则。例如，可以分配一个特定的主机子集用于在商业 Web 服务器上进行可靠的事务处理，而对视频点播的服务则由包含专用硬件的另外一个主机子集来支持。

本发明的一个目的是改进封闭群集的总体通量。

本发明的另一个目的是减少对远程服务请求的总体延迟。

本发明的再一个目的是提供一种用指定节点接替出故障连接路由器的装置，使网络用户感觉不出服务的中断。

根据本发明的第一个方面，封闭群集其特征在于有一个网关节点与数个服务器主机。网关节点（1）将一个封闭群集划分为几个虚拟封闭群集；（2）根据按照一个可配置原则规定的当前服务器负荷度量，将输入连接动态分配在一个虚拟封闭群集中；（3）支持该群集的动态配置。

本发明第二个方面，所提供的一种系统和方法能透明地从一个网关节点的故障恢复过来，以向用户提供不中断的服务。按照该方法，一个群集或虚拟封闭群集中的每个节点均保留一份该网关保留的状态信息子集。当该网关出现故障，状态信息就被传送到备用网关。

在较佳实施例中，封闭群集可以表现为：（1）一个虚拟封闭群集（单一的 IP 地址，为所有远程用户所用）；或者（2）多个虚拟封闭群集（有数个 IP 地址别名）。TCP 连接路由器节点拥有这些 IP 地址，并接收它们所有的 TCP 连接请求。每个 IP 地址对应一个虚拟封闭群集。TCP 连接路由器按与该虚拟封闭群集有关的权重（weight）将新的 TCP 连接分配给每个虚拟封闭群集中的主机。TCP 连接路由器支持的动态配置允许：动态定义各虚拟封闭群集，动态配置与虚拟封闭群集对应的权重，自动或手动管理虚拟封闭群集（增加或撤销主机、服务，等等）。这种解决方法允许动态地配置、增加和撤销服务器主机，同时避免了在网络中服务器名隐藏的问题。

图 1 表示现有技术水平的一个封闭群集系统；

图 2 表示现有技术水平的一个报文交换器（Message Switch）；

图 3 表示本发明一个实施例中的一个虚拟封闭群集系统；

图 4 表示本发明另一个实施例中的一个虚拟封闭群集系统；

图 5 是图 3 和 4 中执行器（Executor）的细节图；

图 6 是图 3 和 4 中管理器（Manager）的细节图；

图 7A-7C 是该执行器的程序流图;

图 8 表示该执行器的数据结构;

图 9 是该管理器的程序流图;

图 10 表示本发明一个实施例中的具有高可用性网关的群集。

本文叙述的虚拟封闭群集系统可作为美国专利 5,371,852 号的一种改进。请参阅 1992 年 10 月 14 日申请的美国专利《使一个计算机群集看起来像计算机网络上的一个主机的方法和设备》，转让给与本发明相同的受让人，申请号是 960,742，专利号是 5,371,852，该专利结合于此作为参考，就像全文复制如下。图 1 是美国专利 5,371,852 号发明所述封闭群集的一个实施例。像美国专利 5,371,852 号的系统一样，本系统为穿越一个计算机群集的边界的 TCP 信息选择路由。该信息是一种端口类型的报文。输入报文被选择路径，服务器作出响应，使得每个群集对外部主机来说就像是一台单独的计算机。本系统中，一个群集要被划分为数个虚拟群集（虚拟封闭群集）。每个虚拟封闭群集在该群集之外的网络上的其它主机看来就像是一台主机。报文被选择路由到每个虚拟封闭群集中的成员，使该组群集节点的负荷保持平衡。

图 3 表示一个用于 TCP 协议族的连接路由器 - TCP 连接路由器（ TCP - CR ） 300 的一个实施例。该装置包含两个以上的计算机节点（ 105 - 109 ），它们由一个被称作互连线路 110 的通信链路连接在一起，形成一个群集。（请注意，在本发明的一个实施例中，互连线路可以是一个网络。）群集中有一个计算机，承担网关 109 的角色，它通过被称为网络 120 的另一条通信链路与一个或多个外部计算机或群集（主机）相连。一个网关可以与多个网络相连，群集中用作网关的节点也可以不止一个。网关与网络的每个连接处（即边界）可以有多个网络地址。每个网关有一个 TCP 连接路由器（ TCP - CR ） 300，如图 10 所示，后者由管理器 320 和执行器 340 组成，还可选用一个恢复管理器（ recovery manager ）。管理器通过向执行器发出命令请求 344 并评估反馈信号 346，对路由的选择进行控制。执行器由一个报文交换器 140 和一个虚拟封闭群集路由器 310 组成，其中报文交换器与美国专利 5,371,852 号所述系统中的类似。

图 4 表示了本发明的另一个实施例。与较佳实施例中一样，群集的节点 107 直接向用户 130 反馈它们的应答（ response ）。然而，本实施例中没有采用图 3 中所示的专用互连线路 110，而是通过外部网络 120 连接所有的群集节点。 TCP 连接路由器也保持不变。样本请求信号 348 从用户 130 经过网关

109，通过外部网络 120 到达群集节点 107。对应该信号的应答 350 直接从节点 107 通过外部网络 120 到达用户 130。

管理器 320 模块执行连接分配政策，允许动态配置虚拟封闭群集。管理器通过动态反馈控制回路对每个封闭群集中成员当前的负荷进行检查和评估。管理器执行的连接分配政策，智能化地将输入连接在虚拟封闭群集的服务器上进行分配，以加快对用户请求的服务。新的权重分配是通过一个管理器算法计算出来的，该算法可由群集管理员进行配置。这种用于权重分配的决策算法的输入参数包括已评估负荷量度以及可由管理员配置的诸如时间阀值之类的参数。输入连接是根据上述输入参数被动态分配到每个虚拟封闭群集的，以保证群集资源的分配为用户提供最快捷的服务。管理器还包括一个命令接口，由管理人员用来动态配置虚拟封闭群集。后面对管理器有更详细的叙述。

假若 TCP 连接路由器节点 109 停止工作，该群集的所有节点就不能向它们的远程用户提供服务。为了解决这个问题，我们增加了一个恢复管理器。当正在工作的网关节点发生故障，恢复管理器就在指定的备用网关节点中启动，帮助服务器节点保留恢复数据。用户无须采取任何行动来从网关故障中恢复过来，而是继续接受群集的不间断服务。后面对恢复管理器有更详细的叙述。

图 5 表示的是执行器 340 的较佳实施例。执行器由命令处理器 540、报文交换器 140 和虚拟封闭群集路由器 310 组成。命令处理器 540 为执行器 340 接收请求并返回应答 346。命令处理器与报文交换器 140 和虚拟封闭群集路由器 310 交互作用，执行请求，产生应答。命令处理器可以修改连接表 510、虚拟封闭群集表 550、端口表 520 或服务器表 530 的内容。报文交换器 140 和连接表 510 与美国专利 5,371,852 号中的报文交换器和连接表相同。在本发明的较佳实施例中，虚拟封闭群集路由器 310 不改变输入包。包被传送到服务器，对服务器所作的配置使得应答从内部节点直接被发送到用户。

报文交换器 140 基本上与美国专利 5,371,852 号中的报文交换器相同。不过，本发明对较佳实施例中的报文交换器进行了优化，在报文交换器中增加了一个检步骤（ check ）。报文交换器必须检查报文的发送对象是否是虚拟封闭群集路由器已知的一个虚拟封闭群集。

虚拟封闭群集路由器中保存一组对外部网络上用户来说代表各个虚拟封闭群集的地址。虚拟封闭群集路由器向群集的内部节点传送请求，并不修改

所接收的请求。群集的每个内部节点与一个或多个虚拟封闭群集相关联，并且只接收对其所关联的虚拟封闭群集所作的请求。运用现有水平的技术，本发明中的内部节点被配置为能接收发向代表一个虚拟封闭群集的地址的包，且能直接向用户作出应答。现有技术要求报文交换器 140 必须改写输入请求的包的首标（图 1 中 140）和对请求的应答的包的首标（图 1 中 120）。本发明中，不必改写包的首标。（现行技术可以用于本发明中。）本发明的性能优于现行技术，因为包的首标不作改写，应答包的传递不经过网关节点 109。因为应答包的传递不经过 TCP 连接路由器，报文交换器就不接收来自群集内部节点的应答包。其结果是，在较佳实施例中，省略了对首标的改写，省略了对来自内部节点的应答包的检查。

这种改进的一个直接后果是，虚拟封闭群集路由器只能看到在用户与提供服务的内部节点之间传递的数据流的一半。这就对保持连接表的准确造成了困难。为解决这个问题，本发明使用了连接表专用的两个新定时器（timer）：一个失效超时定时器（stale timeout），一个结束超时定时器（FIN timeout）。用了这两个定时器，加上通信流和现有技术水平已知的定时器，就能准确地保持连接表。

连接表内的条目有两种状态：活动（ACTIVE）状态和结束（FIN）状态。每当建立了一个新连接，就在连接表中增加一个条目，将其置为活动状态。每当一个包经过对应连接表中某条目的连接时，就在该连接条目上打上时间标记。当虚拟封闭群集路由器看到一个结束信号（FIN）从用户流到提供服务的节点，相关联的连接表条目就被置为 FIN（结束）状态。（包可以继续在被置于 FIN 状态的连接上传递。）连接表中的一个条目被关闭并等待被清除的时机是，自用户在该连接上向服务器发送上一个包开始后的时间，超过了结束超时定时器确定的时间。如果用户出错，没有发送结束信号，该连接记录条目不变。失效超时定时器规定的是，在上一个包在活动对话（conversation）中被传递后，清除该连接表条目之前的等待时间。

图 7A - 7C 表示的是虚拟封闭群集路由器 310 的流程图。图 7A 中，在步骤 702，虚拟封闭群集路由器等待一个包。当包被接收后，在步骤 704，虚拟封闭群集路由器检查该包是用于一个已有的 TCP 连接还是一个新的 TCP 连接。如果该包是要用于一个已有的 TCP 连接，则在步骤 708 检查该包的类型是结束（FIN）、同步（SYN）还是重设（RST）（都是现有技术已知的包类型）。如果该包不是上述其中之一，则在步骤 722，虚拟封闭群集路

由器就将该包传递到该连接所关联的内部节点。否则，在步骤 710，检查该包是否是个 RST（重设）。如果该包是个重设，则在步骤 712 该对话被从连接表中清除，重新设定（resetting）连接，然后在步骤 722 将该包传递到与该连接相关的内部节点。如果该包不是个重设，虚拟封闭群集路由器 310 在步骤 714 检查该包是否是同步。如果该包是 SYN，就在步骤 716 建立连接，即使该连接此前存在，也将该连接置于活动状态。然后虚拟封闭群集路由器 310 在步骤 718 检查该包是否是 FIN（结束）。如果该包是 FIN，则在步骤 720 该连接被置于 FIN 状态。经有关结束状态处理之后，或者如果该包不是 FIN，该包在步骤 722 被传递到该连接关联的服务器。

图 7B 表示连接不存在情况下的流程图。如果在步骤 704 的检查发现该包是要经过一个非现有的连接，则在步骤 724 虚拟封闭群集路由器首先检查该包是否是 SYN（同步信号）。如果该包不是 SYN，则在步骤 726 该包就被放弃。如果该包是 SYN，则在步骤 728 建立一个连接并置其于活动状态 728，在步骤 730 选择一个服务器，而在步骤 722 将该包传递到所选择的服务器。

图 7C 表示的是为一个新连接选择服务器的步骤 730 的过程的流程图。本发明中，本功能按权重进行路由选择。为叙述服务器选择方便起见，我们把虚拟封闭群集的内部节点从 1 至 n 编号。例如，如果一个虚拟封闭群集有七个节点，则将它们分别编号为 1、2、3、4、5、6 和 7。为叙述服务器选择方便起见，我们还把合格权重从最大有效值至 1 进行编号。例如，设最大有效值为 5，则各合格权重分别为 5、4、3、2 和 1。零是一个特殊值。对合格权重的选择按降序进行。本发明为提供特定服务的每个内部节点赋予一个权重。保证对各个服务，要么至少有一个节点有非零的最大权重，要么所有节点的权重均为零。

选择服务器的过程 730 首先在步骤 734 读取次高级服务器对应的编号和当前的合格权重。然后在 735 检查这个编号是否太大。如果该编号不是太大，就在步骤 746 检查与该编号对应的服务器是否合适（后文将叙述该检查过程）。如果该编号太大，则在步骤 736 选取具有下一具有较低权重的第一个服务器，然后在步骤 738 检查该下一较低权重是否等于零。如果该下一较低权重不等于零，就以此来作为当前合格权重，该过程继续在步骤 746 检查当前选择的服务器是否合适。选择了最大权重的第一个服务器后，该过程在 724 检查是否有任何可用的服务器可作为包传递的目的地。如果所有可用节点的权重均为零，则所有的服务器均不能提供服务。如果没有能提供服务的服务器，

就不选择服务器，该包在744被返回。如果有服务器可用，该过程在步骤746检查是否有一个好的选择。所谓好的选择，是指所选择的服务器，其权重大于或等于当前合格权重。如果有服务器是好的选择，就在748选择该服务器，并在步骤750将其返回给虚拟封闭群集路由器。如果这不是好的选择，该算法就在步骤734挑选下一个服务器。

因为要么最大权重不等于零且至少有一个节点具有最大权重，要么所有节点的权重均为零，所以选择服务器的过程总会要结束。当有权重大于零的节点时，选择服务器的过程按照各权重的比例分配包。例如，对于权重分别为3和2的两个内部节点，权重为2的节点每接收2个包，权重为3的节点就接收3个包。

图8是一例虚拟封闭群集路由器所用数据结构。虚拟封闭群集表550含有在外部网络上的虚拟封闭群集地址的地址集。该表还包括与虚拟封闭群集具体关联的所有参数。每个虚拟封闭群集与一个端口表520关联，端口表含有该虚拟封闭群集为之提供服务的所有端口。每个端口项802对应一个失效超时定时器804、FIN（结束）超时定时器806以及其它端口描述属性808.每个端口与一组用于提供与该端口关联的服务的虚拟封闭群集内部节点的子集相关联。节点表530含有与该端口关联的各节点820的地址、与该节点关联的当前权重822以及其它节点特点数据830。（节点特点数据的一个例子是计数器，分别用于指示处于活动状态连接的数目、处于FIN（结束）状态连接的数目以及已完成连接的总数。）节点表530还含有选择服务器的过程所需的状态信息，以便其对该表中所有节点进行加权路由选择。节点表所含数据还有：节点810的总数、上一个被选节点812、当前合格权重814、最大权重816以及权重界限818。权重界限用于限制最大权重的变化范围。节点具有的权重不得大于权重界限。

本发明的连接路由器管理器（管理器320）的方法和设备，按照可配置的政策用几种负荷量度（load metrics）来动态分配输入连接。管理器中有一个控制回路，动态地修改执行器340路由选择算法的权重，以使对群集资源的分配最优化。本发明的目标是，根据群集当前的状态分配输入TCP连接，以提高群集的总体吞吐量，减少对服务请求的总体延迟。为此，本发明描述了一种向服务器主机分配连接的方法，以提高服务器的使用率，减少为请求服务的时间延迟。

图6所示作为本发明管理器320的一个实施例，其所在的群集600有5个

节点（分别为 105、106、107、108 和 109）。图 6 中的网络配置方案是基于图 4 所示方案的一个替代方案，当然，图 3 所示的配置方案也是可行的。其中一个节点是网关 109，它与外部网络 120 相连，行使 TCP 连接路由器 300（执行器 340 和管理器 320）的功能。管理器 320 由 5 个基本部分（generic components）组成：负荷管理器（Mbuddy）610、外部控制接口（Callbuddy）620、群集主机量度管理器（Hostmonitor）630、正向量度生成器（Forward Metric Generator - FMG）640 和用户可编程量度管理器（UPMM）650。

负荷管理器 610 可用四个不同类别的量度为执行器计算权重函数：输入量度、主机量度、服务量度以及用户量度。负荷管理器 610 接收的这些量度及其它有关数据来自执行器接口 346、外部控制接口的接口 624、群集主机量度管理器接口 634、正向量度生成器接口 644 和用户可编程量度管理器接口 654。负荷管理器 610 通过接口 344 控制与执行器路由选择算法关联的每个虚拟封闭群集端口服务器的权重。

负荷管理器 610 周期地通过接口 346，向执行器 340 请求得到关于每个服务器的各内部计数器的数值。例如，它周期性地请求得到为每个服务器所接通连接的总数的各计数器的数值。负荷管理器 610 能通过将对在 T1 和 T2 两个不同查询时刻的该服务器的计数器内容相减，对一个量度变量进行计算，该变量表示在时间段 T1 - T2 期间所接收的连接的次数。所有这样计算出的输入量度的集合近似地描述了对每个虚拟封闭群集和每个端口服务连接请求的特征率。

群集主机量度管理器 630 通过报文接口 634，周期地向负荷管理器 610 提供群集中每个主机的状态信息。获得这种状态信息的方法有许多。例如，群集主机量度管理器可以用监控中介（monitoring agents）635 执行各种程序脚本来评估主机特定的量度。例如，某个脚本评估的是当前为网络连接对存储缓冲器的利用程度。如果在一个政策特定的阈值时间内未收到关于某个主机量度的报告，就赋予该量度一个特殊的值，管理器就可以判定主机不能再提供服务，于是就不再向其传送连接请求。群集主机量度管理器 630 负责收集整理来自所有监控中介的报告，将其传送给负荷管理器。

正向量度生成器 640 用正向请求生成应用特定或服务特定的量度，并对它们进行评估。正向请求就是在网关计算机 109 上产生的请求。评估过程包括，对每个群集主机服务器产生适当的请求，并检测各服务器的响应延迟时间。

例如，要得到一个 HTTP 服务器上的正向延迟量度，正向量度生成器可对服务于某特定端口（如端口 80）的群集中的每个 HTTP 服务器生成一个 HTTP “GET/” 请求。然后，正向量度生成器 640 检测各个服务器为 HTTP 请求服务的相应的延迟时间，并向负荷管理器 610 传送含有各量度数据的向量。如果在政策特定的阈值时间内该请求没有被应答，正向量度生成器 640 就将与其对应的服务节点标示为暂时不接受该特定类型的新的服务请求。管理器用这个信息来确定暂时不能提供服务的主机，于是就不再向其传送该类型的连接请求。

用户可编程量度管理器 650 允许本发明的用户定义新的任意量度，对连接进行管理。新定义的量度可以定义要在任何特定群集执行的任意规则。例如，任意规则可以要求，出于管理上的原因，某一组群集主机在某些时间段内不得接受任何 TCP 连接。用户可编程量度管理器 650 通过接口 654 将表现为量度的这些规则传达给负荷管理器。

外部控制接口 620 允许管理人员动态调整负荷管理器 610 的任何参数。外部控制接口允许管理人员配置该算法，计算由负荷管理器执行的权重的分配。例如，管理人员可能想要动态改变与当前每个量度相关的权重。管理人员可以选择的方案例如有：（1）提高主机量度的权重；（2）降低服务量度的权重；（3）提高向执行器 340 查询输入量度的频率。外部控制接口 620 通过接口 622 接受管理人员的请求，通过接口 624 传达给负荷管理器 610。

负荷管理器 610 在各服务器与连接路由器网关节点之间建立一个动态反馈控制环路。负荷管理器对执行器 610 路由选择算法中的权重进行调整，以便使根据负荷量度标准负荷较轻的服务器，接受更多的适合其类型的 TCP 连接请求。如果定义了如上所述的一组负荷量度和规则量度，负荷管理器就会根据每个虚拟封闭群集中每个端口的每个服务器的当前量度和当前权重，为之计算新的相对权重。

计算每个虚拟封闭群集中每个端口的权重分配的方法如下：（1）计算所有正在执行的服务器的总量度（AM）；（2）计算每个正在执行的服务器的当前的权重比例（CWP）；（2）对每个量度 M 为每个服务器 S 计算其值（相对于总量度 AM）的量度比率（MP）；（3）计算每个服务器的新权重（NW）：3a）如果服务器处于静态，置 NW 为 0；3b）如果服务器具有固定权重 W，NW 的值为 W；3c）计算向量 NWV，其中，向量的每一项 NWV[i]是一个量度 M[i]的函数，具体公式为 NWV[i]=AW + [(CWP -

MP)/SF]，其中 AW 是当前权重范围中的平均权重， SF 是一个平滑系数参数； 4d ）按下列公式计算每个服务器的新权重:

NW = NWV[1]*w[1] + NWV[2]*w[2] + … + NWV[i]*w[i]

图 9 的流程图，描述了管理器如何接收量度以及本发明如何计算权重分配。方框 910 表示负荷管理器处于等待状态，等待的事件要么是一个报文，要么是一个超时信号（ timeout ）。判断框 920 进行判断，确定发生的事件的类型。如果事件是超时信号，要求进行数值刷新，则至方框 930，询问执行器，要求提供存放输入量度的一组计数器的值（方框 935 ）。如果判断框 920 确定事件是请求进行参数更新，则在方框 928 对相应的参数进行更新。例如，管理人员可以对与任何量度关联的权重、或对查询的时间段进行更新。如果判断框 920 确定，事件是接收量度更新，则至方框 925，读出量度，并相应地设置各内部变量的值。如果来了新量度，方框 940 表示的算法就计算所有量度的当前比例和当前的各个权重。然后至方框 950，为每个服务器节点计算新的权重 NW，计算公式如上所述。该计算的结果是一个权重 NW[i] 向量，其中每项 NW[i]代表服务器 i 的新权重。判断框 960 进行判断，用一个任意阈值函数（ arbitrary threshold function ）确定计算出的新权重向量 NW[i]是否与当前的权重向量不同。如果不同，至方框 970 ，将新权重传给执行器；否则，程序返回至顶部方框，等待新的事件。

当检测到正在工作的网关发生故障时，指定的后备网关中的恢复管理器启动。关于故障检测的一般方法的代表论文有： A Bhide 等人的《高利用率的网络文件服务器》（见 199 页，该文 1991 年冬在美国德克萨斯州达拉斯举行的 USENIX 大会上发表，原文名为 A highly available Network File Server ）; F. Jahanian 等人的《处理器组成员协议：说明、设计与实现》（见 2 - 11 页，该文 1993 年十月在新泽西州普林斯顿 IEEE 计算机协会举办的第 12 届可靠分布式系统研讨会上发表，原文名为 Processor Group Membership Protocols: Specification,Design and Implementation ）。

按照《高利用率的网络文件服务器》中所述方法，恢复管理器首先从出故障网关中撤销网络连接，然后查询所有活动的服务器节点，从它们的影子连接表（ shadow connection table ）中获得状态信息，并根据该信息在网关的报文交换器中构造连接表。接替过程必须在 TCP/IP 的超时间隔（ timeout interval ）内完成，以免丢失已有的连接。为此，内部节点执行一个新的混合算法（见下文），检测何时连接变为不活动状态，将这些连接从影子连接表

中去除，从而只将处于活动状态的连接向接替网关进行说明。当所有正在工作的群集节点都作出响应（在一定时间内不作出响应的节点被视为不起作用的节点），备用网关中的恢复管理器执行器，激活其自有的网络接口，以便接收向群集 ip-地址寻址的包。这最后一步，就完成了让备用网关运行所需的工作。管理器使用的相对静态配置数据被保存在一个由主网关和备用网关共享的文件，备用网关在接替过程中读取该数据。

一种明显的替代解决方案是，在备用网关中双份保存连接数据，但该方案并不可取。它要求在主网关与备用网关之间关于每个接通连接和终止连接有一个“两阶段”协议，由于费用太高，所以不能采用。

图 10 示出具有高可利用率网关的一种封闭群集的配置。主网关 1050 被连接到外部网络 120，连接处于活动状态。指定的备用网关 1030 物理连接到网络 120，处于不活动状态。除了一般封闭群集网关部件管理器 320 和执行器 340 之外，每个网关还有一个恢复管理器 1020。（主网关在出现故障、恢复处理后，可以变成备用网关。）每个服务器节点 107 含有一个影子连接表 1010，里面保存着有关其与外部网络 120 的处于活动状态的连接的信息。

报文（ip 包）到达群集网关，被导向到特定的 TCP 或 UDP 协议端口。网关中的报文交换器允许为协议端口安装一个报文路由选择功能。对每个到达关联端口的报文，路由选择功能被调用一次，负责选择报文要被传至的内部节点和端口。表示已接通通信连接以及占用该通信连接的群集节点的数据，被记录在网关中一个表中。报文交换器利用此表，在已接通通信连接上为输入包选择路由，送至正确的群集节点。

相对静态信息，例如哪些服务器端口安装了报文交换器功能等信息，以及其它管理器配置数据被保存在一个共享文件中，可由主网关和备用网关访问。当前连接信息的变化很快，要按照本说明书中的技术进行管理。

每个内部节点 107 保存一个网关路由选择表中的影子连接表 1010，内含该节点自己的连接信息（而不含其它节点的连接信息）。在网关接替期间，节点使用该阴影表，对来自备用网关 1030 中恢复管理器 1020 发出的接替网关请求作出响应。该表大大地节省了内部节点对接替网关请求的响应所需的时间，这一点非常重要，这是因为，为保持已接通通信连接不中断，接替网关必须在基于连接的协议在成功地完成一次通信所允许的“超时”期限内工作。

为了在网关以及内部节点中保存的影子表开辟空间，存放连接表中的条

目，我们采取以下方法。连接的状态要么是活动（ Active ），要么是结束（ FIN ）。连接表中每个条目在每一引用上均有一个时间标记。保存一个用户可配置的定时器 FIN_TIME_OUT 。该定时器表示最后一次引用处于结束状态的对话后，该对话将被关闭的时刻。定时器可以是通用的，也可以是服务地址专有的或端口专有的。主动关闭的目的（主动关闭指通信连接的一端发出结束信号（ FIN ）而另一端继续在该通信连接上发送信息）是，允许服务器继续向用户发送数据，数据发送完毕，对话就被关闭。用户被允许主动关闭对话，以向服务器表示将不传送来自该用户的进一步的请求。为叙述方便，我们假设用户的请求是通过路由器发送的。这种协议之所以行之有效，是因为服务器继续发送得到确认的数据。路由器以及服务器接到确认信号，不断地在连接表条目中作上时间标记。一旦服务器结束向用户发送数据，关闭它的那一半对话，服务器将得到来自用户的最后一次确认信号。定时器 FIN_TIME_OUT 规定的时间过后，服务器就能清除该连接条目。第二个定时器 STALE_TIME_OUT 在网关中。如果处于活动状态的连接超过 STALE_TIME_OUT 规定的时间而不发生通信动作，则该连接可以被清除。

内部节点中也执行该算法（连接重建算法），为内部节点保存的、支持备用网关的接替过程的连接表的影子表中开辟条目位置。用这种方法，可以使影子表中的条目的数量尽可能地少，从而使得接替过程尽可能快地进行。

FIN_TIME_OUT 的缺省值被置为 TCP 的最小段长（ MSL)的三倍。 STALE_TIME_OUT 的缺省值应该大于 TCP 的失效超时值（ stale time out ）。要定出 FIN_TIME_OUT 的更合理的值，就要对与该定时器关联的协议加以具体的考虑。

当由于发现主网关发生故障（由非本公开的一部分的在（ 5 ）中叙述的类型外的一些装置确定）或是出于管理需要而决定要启用备用网关作为群集的网关时，备用网关 1030 中的恢复管理器 1020 采取以下步骤:

（ 1 ）备用网关用步骤（ 3 ）中所述的 ip 地址接替方法，取消主网关的网络连接。该步骤旨在保证被认为替换下来的网关不能从网络上接收数据。有一种类型的故障（网关发生部分故障），要是不采取该步骤，故障网关就有可能继续接收报文并进行处理，这是对系统完整性的折衷。

（ 2 ）备用网关询问群集的每个作用节点，要求提供分配在各节点的所有 UDP 端口的描述数据，以及在其与群集外主机之间通过主网关建立的 TCP 连接的描述数据。这由备用网关用一种基于 ip 的专用协议来作。每个节点中

的影子连接表允许节点立即作出响应，提高了已建立的连接不在网关接替期间超时的概率 。上述识别关闭连接、要求为支持它们开辟条目位置的算法，使影子连接表最小化，有助于减少完成网关接替所需的时间。

（3）备用网关记录每个操作节点发出的响应信号，将节点的 UDP 端口和 TCP 连接记录在备用网关的执行器 340（图 10）中的连接表 510（图 5）中。

（4）当所有工作的群集节点发出响应信号（在规定时间内没有响应的节点被假定为不在工作），备用网关启用自己的网络接口，接收寻址到该群集 ip 地址的包。该最后步骤完成了允许备用网关执行功能所需的工作。

以上通过较佳实施例对本发明进行了描述。熟悉该领域的人可以对其作出各种修改和改进。因此应该明白，本文所述的较佳实施例仅具有示范意义，而不是限制性的。本发明的范围由下文中各权利要求确定。

说 明 书 附 图

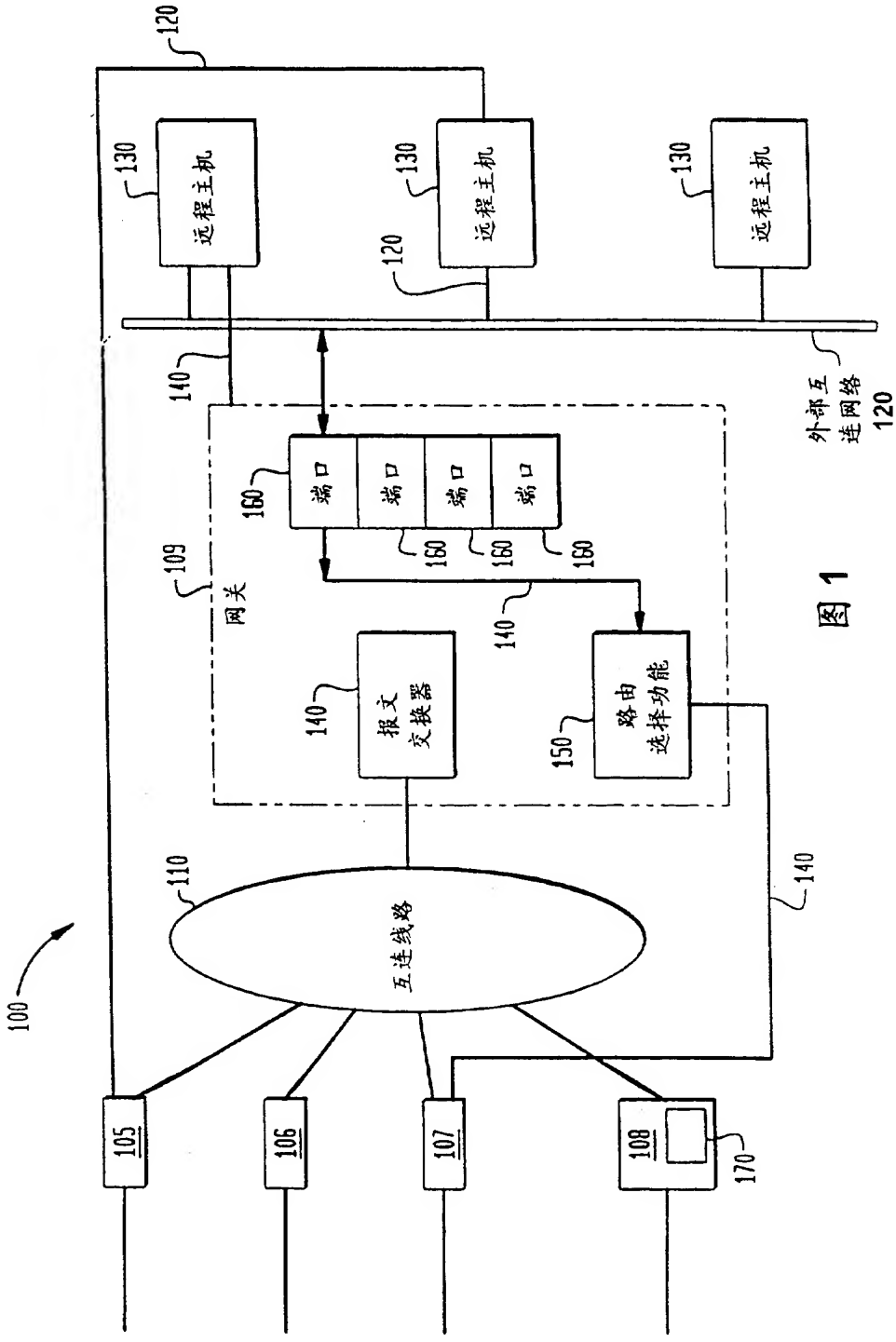

120
130
远程主机
130
远程主机
130
远程主机
120
140
外部互连网络
120
160
端口
端口
端口
端口
160
160
160
109
网关
140
140
报文交换器
150
路由选择功能
140
110
互连线路
100
105
106
107
108
170

图 1

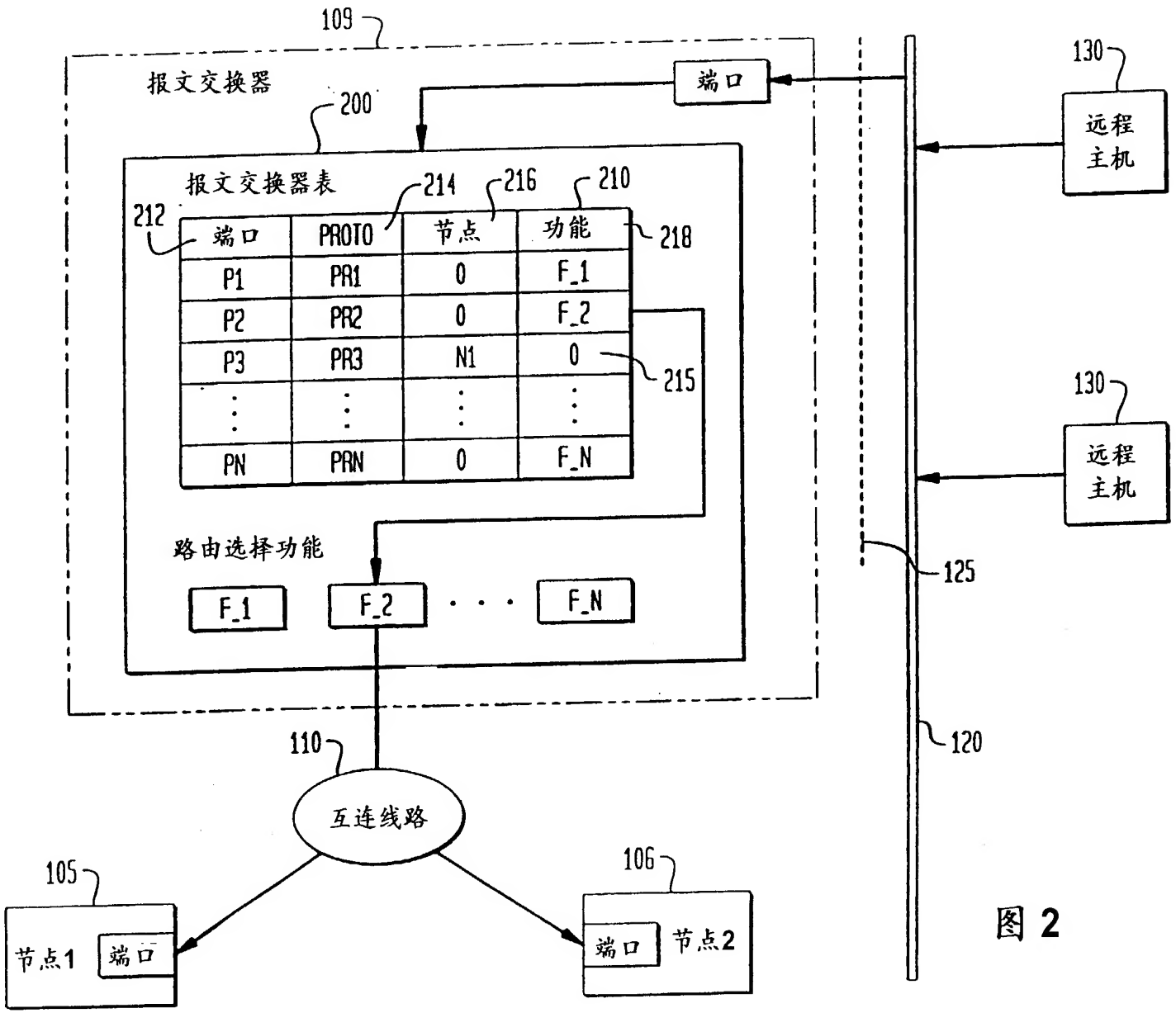
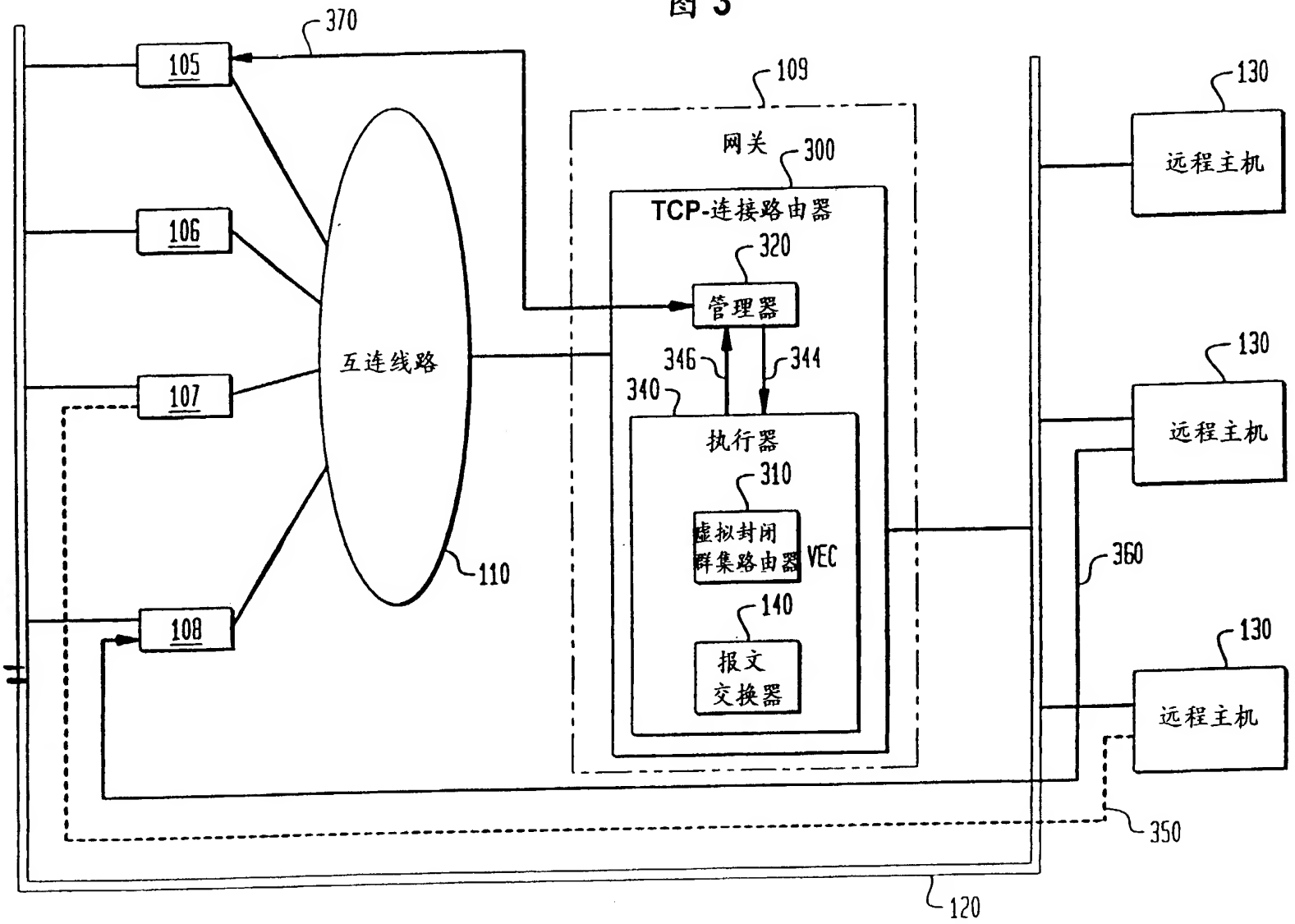

109
报文交换器
端口
200
报文交换器表
214
216
210
212
端口
PROTO
节点
功能
218
P1
PR1
0
F_1
P2
PR2
0
F_2
P3
PR3
N1
0
215
PN
PRN
0
F_N
路由选择功能
F_1
F_2
F_N
130
远程主机
130
远程主机
125
120
110
互连线路
105
节点1
端口
106
端口
节点2

图 2

图 3


370
105
106
107
108
互连线路
110
109
网关
300
TCP-连接路由器
320
管理器
346
344
340
执行器
310
虚拟封闭群集路由器
VEC
140
报文交换器
130
远程主机
130
远程主机
130
远程主机
360
350
120

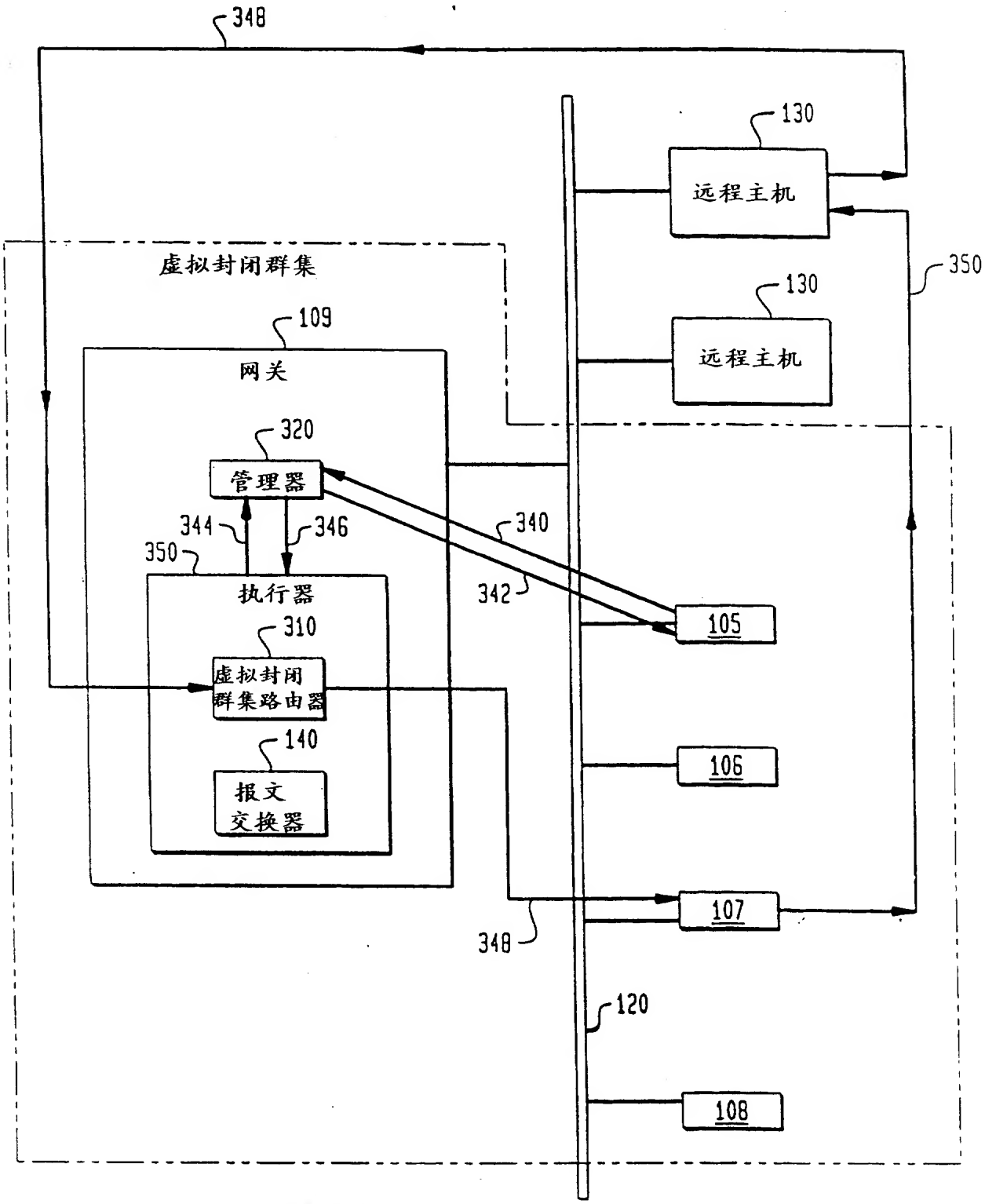

348
130
远程主机
虚拟封闭群集
350
130
远程主机
109
网关
320
管理器
344
346
340
350
执行器
342
105
310
虚拟封闭
群集路由器
140
报文
交换器
106
107
348
120
108

图 4

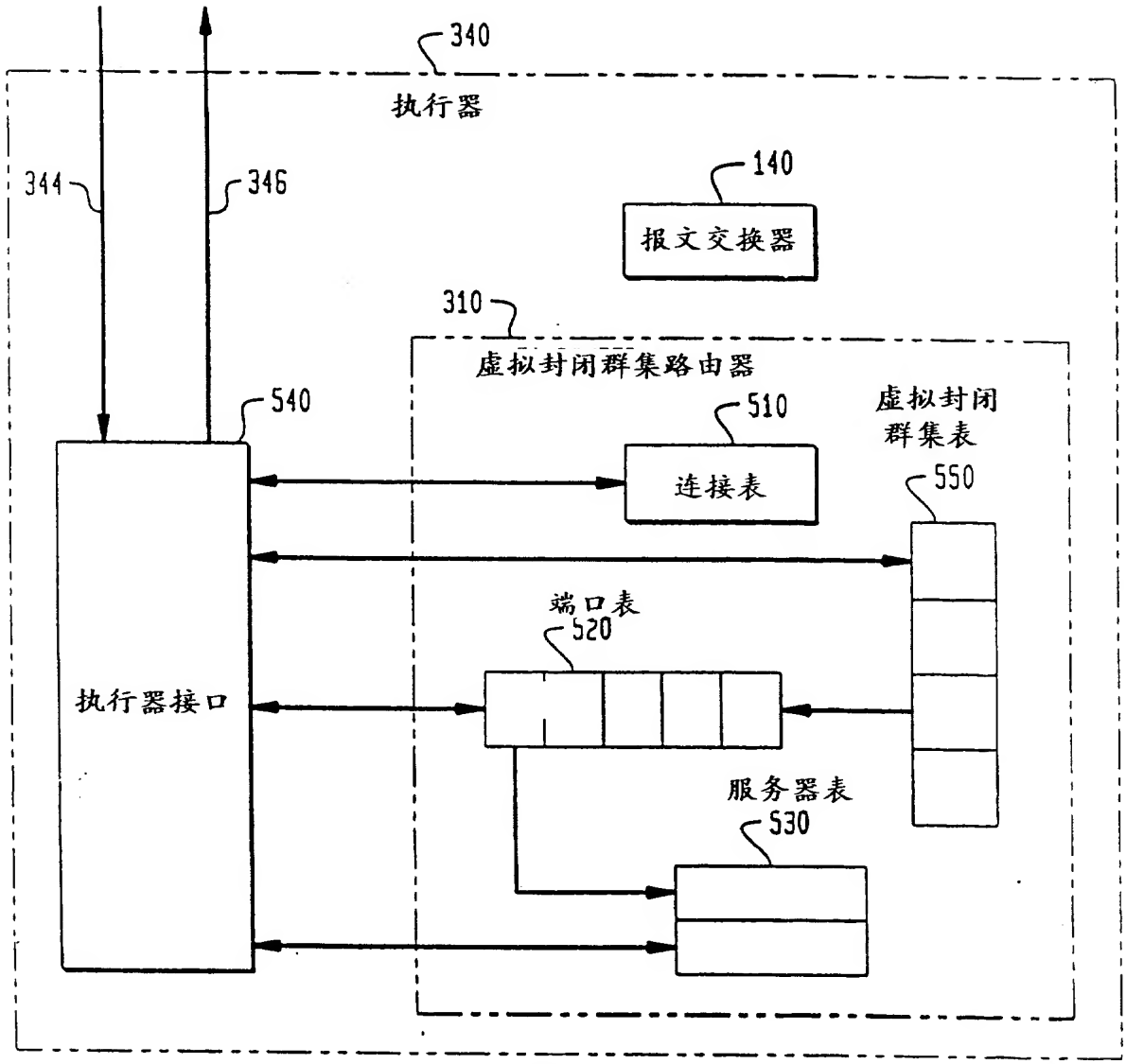

340
执行器
344
346
140
报文交换器
310
虚拟封闭群集路由器
540
510
虚拟封闭
群集表
连接表
550
端口表
520
执行器接口
服务器表
530

图 5

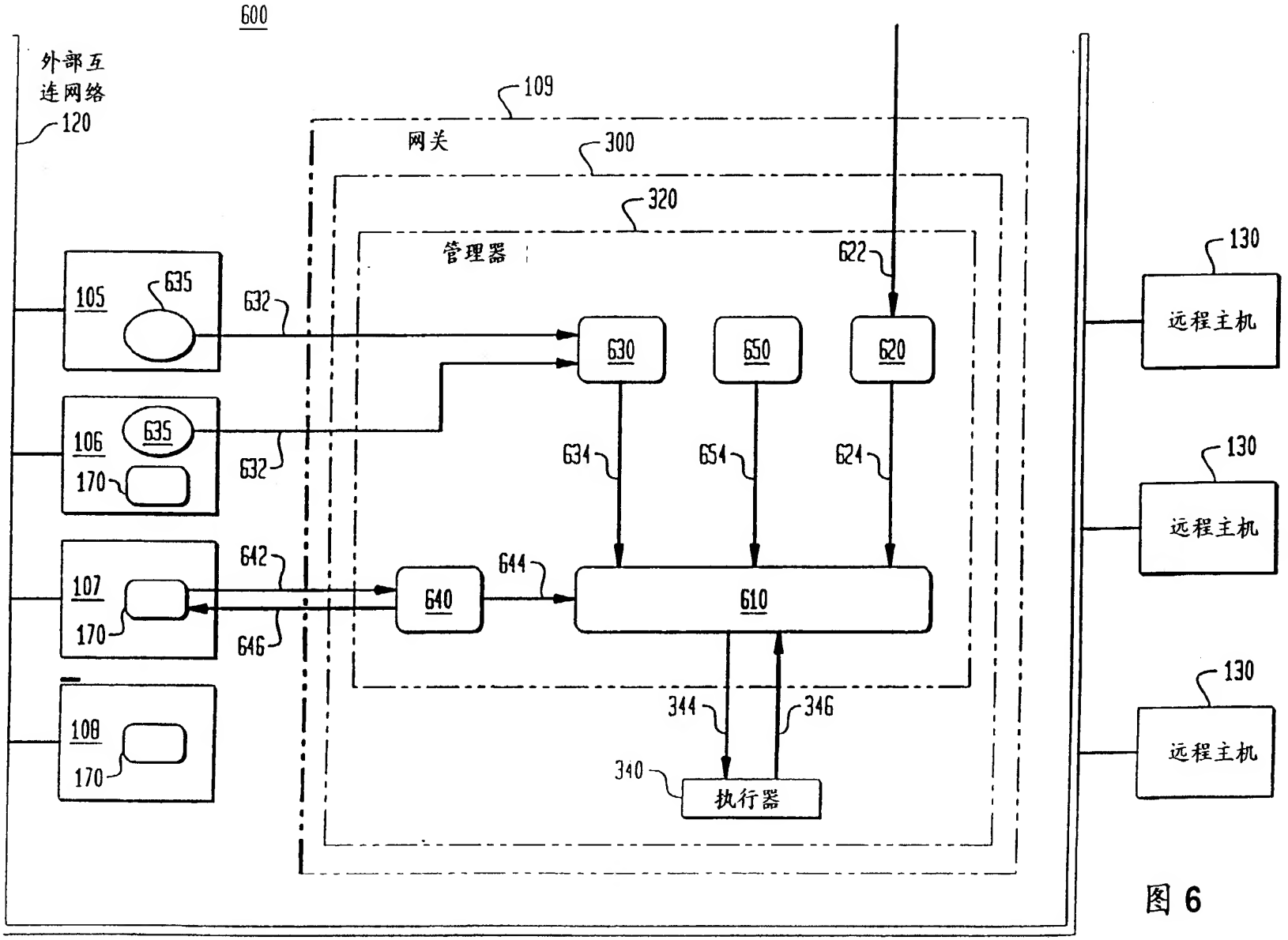

600
外部互连网络
120
109
网关
300
320
管理器
105
635
632
630
650
620
622
130
远程主机
106
635
170
634
654
624
107
170
642
640
644
610
646
108
170
344
346
340
执行器

图 6

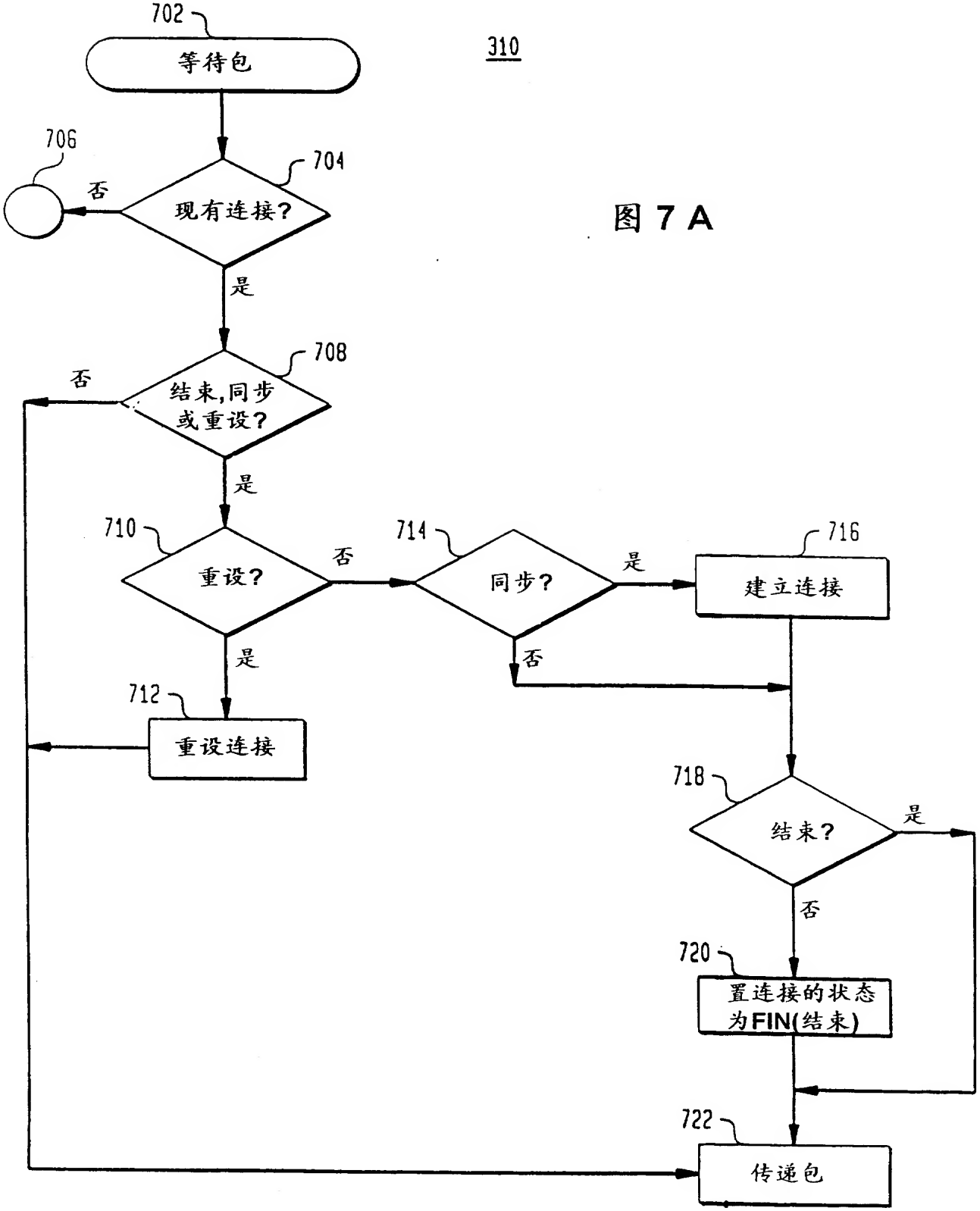

310
702
等待包
706
704
否
现有连接?
是
708
否
结束,同步
或重设?
是
710
重设?
否
714
同步?
是
716
建立连接
否
是
712
重设连接
718
结束?
是
否
720
置连接的状态
为FIN(结束)
722
传递包

图 7 A

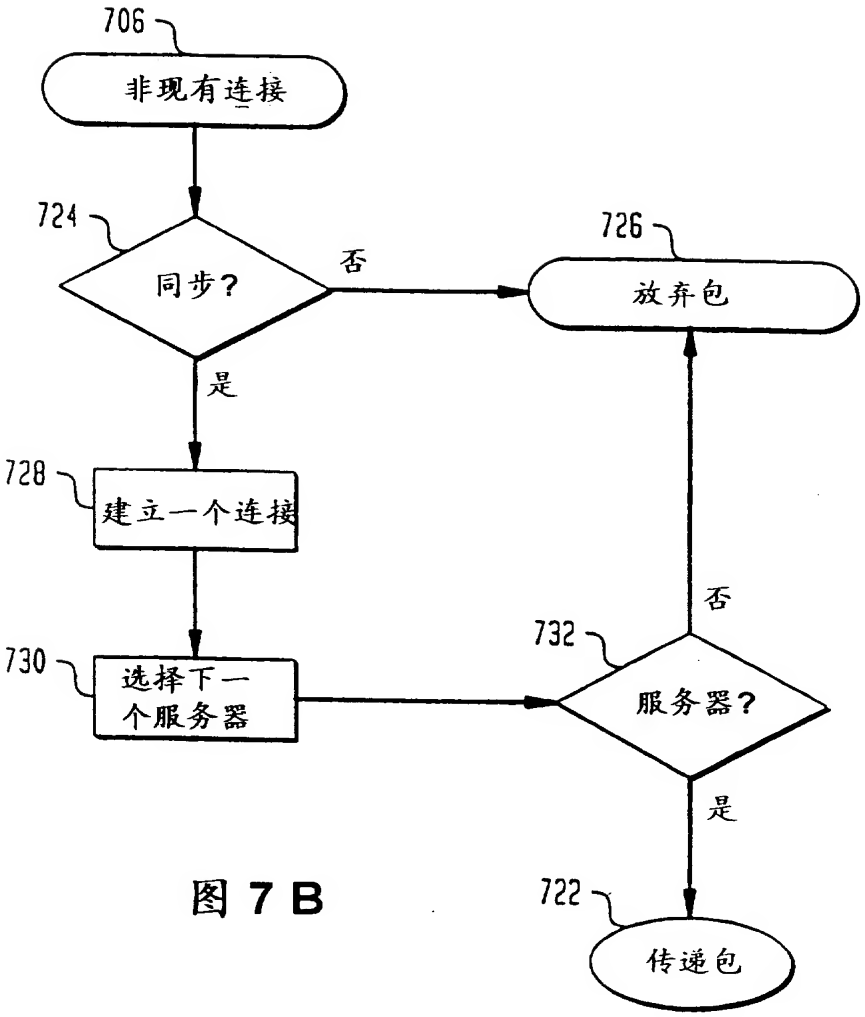

706
非现有连接
724
同步?
否
726
放弃包
是
728
建立一个连接
730
选择下一个服务器
732
服务器?
否
是
722
传递包

图 7 B

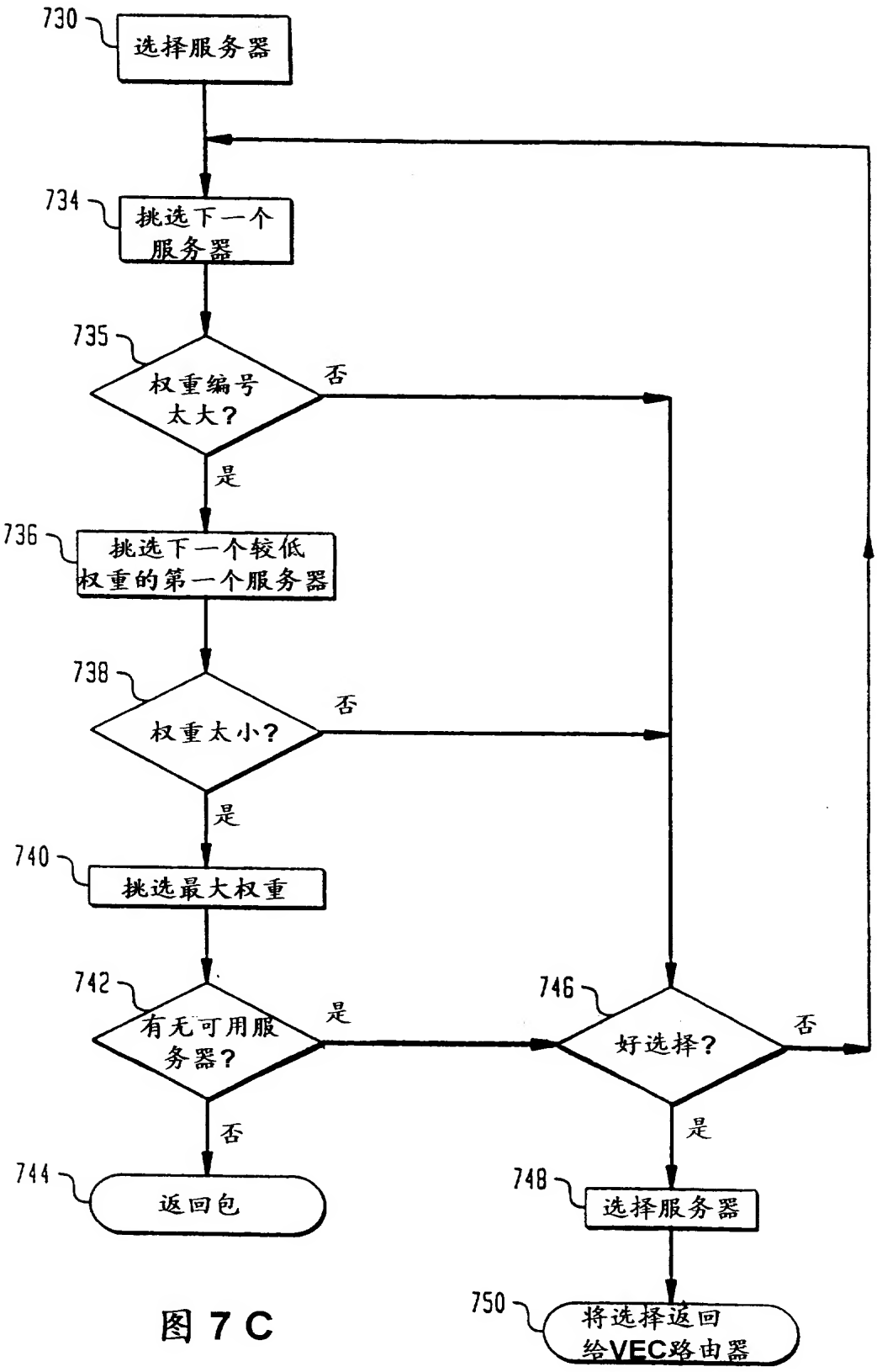

730
选择服务器
734
挑选下一个服务器
735
权重编号太大?
否
是
736
挑选下一个较低权重的第一个服务器
738
权重太小?
否
是
740
挑选最大权重
742
有无可用服务器?
是
否
744
返回包
746
好选择?
否
是
748
选择服务器
750
将选择返回给VEC路由器

图 7C

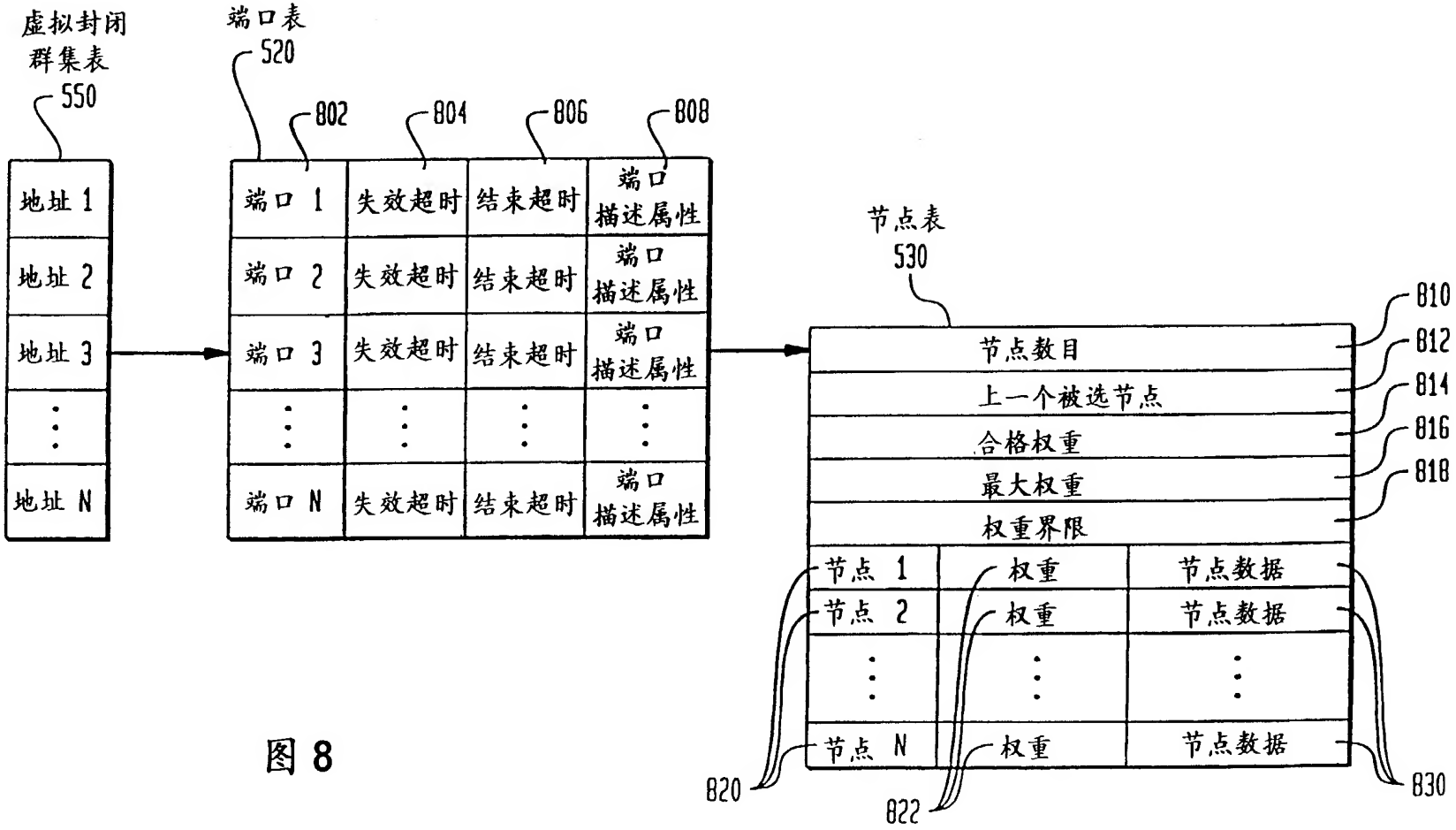

虚拟封闭群集表
550
端口表
520
802
804
806
808
地址 1
地址 2
地址 3
地址 N
端口 1
端口 2
端口 3
端口 N
失效超时
结束超时
端口描述属性
节点表
530
节点数目
上一个被选节点
合格权重
最大权重
权重界限
节点 1
节点 2
节点 N
权重
节点数据
810
812
814
816
818
820
822
830

图 8

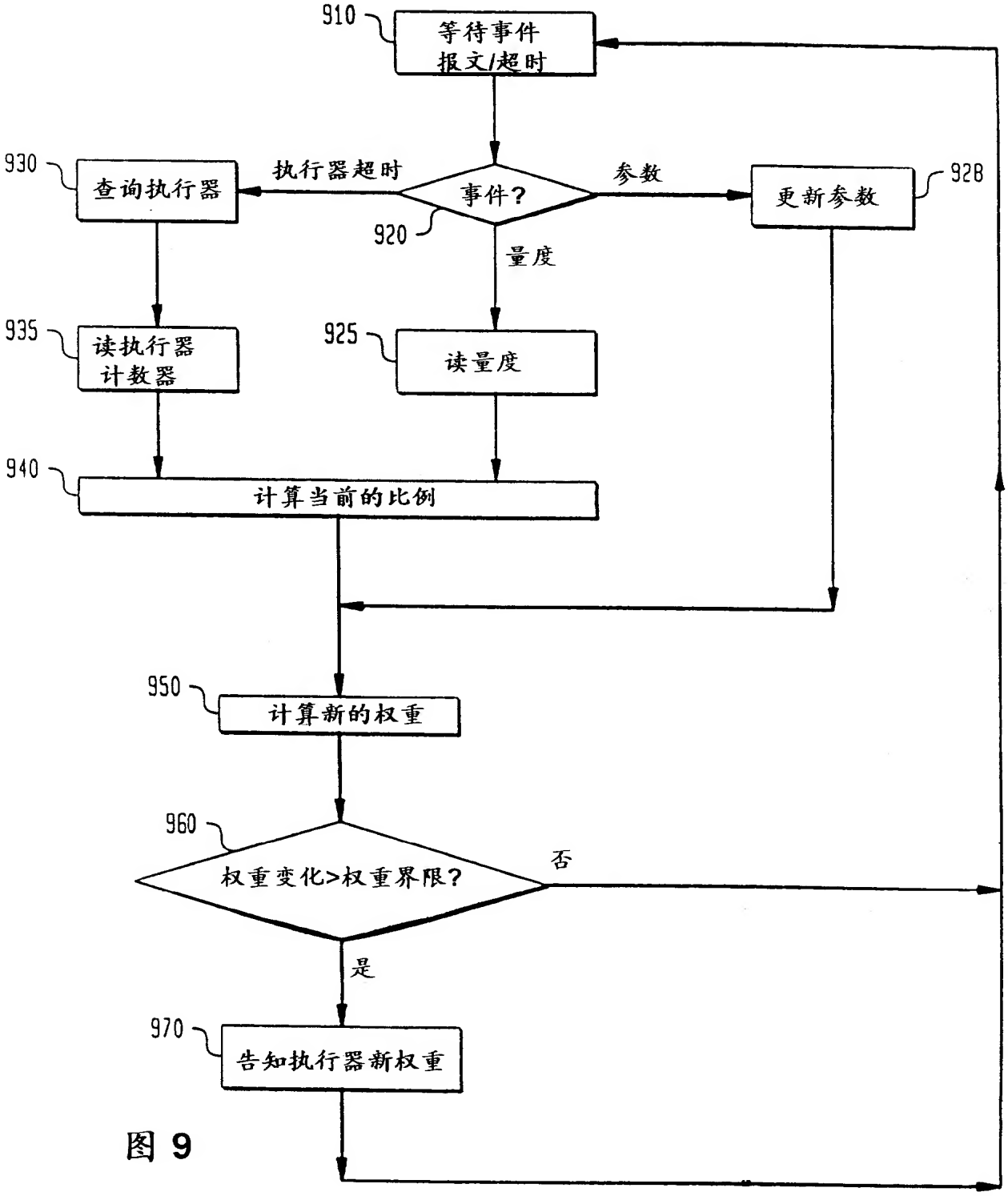

910
等待事件
报文/超时
930
查询执行器
执行器超时
事件?
参数
928
更新参数
920
量度
935
读执行器
计数器
925
读量度
940
计算当前的比例
950
计算新的权重
960
权重变化>权重界限?
否
是
970
告知执行器新权重

图 9

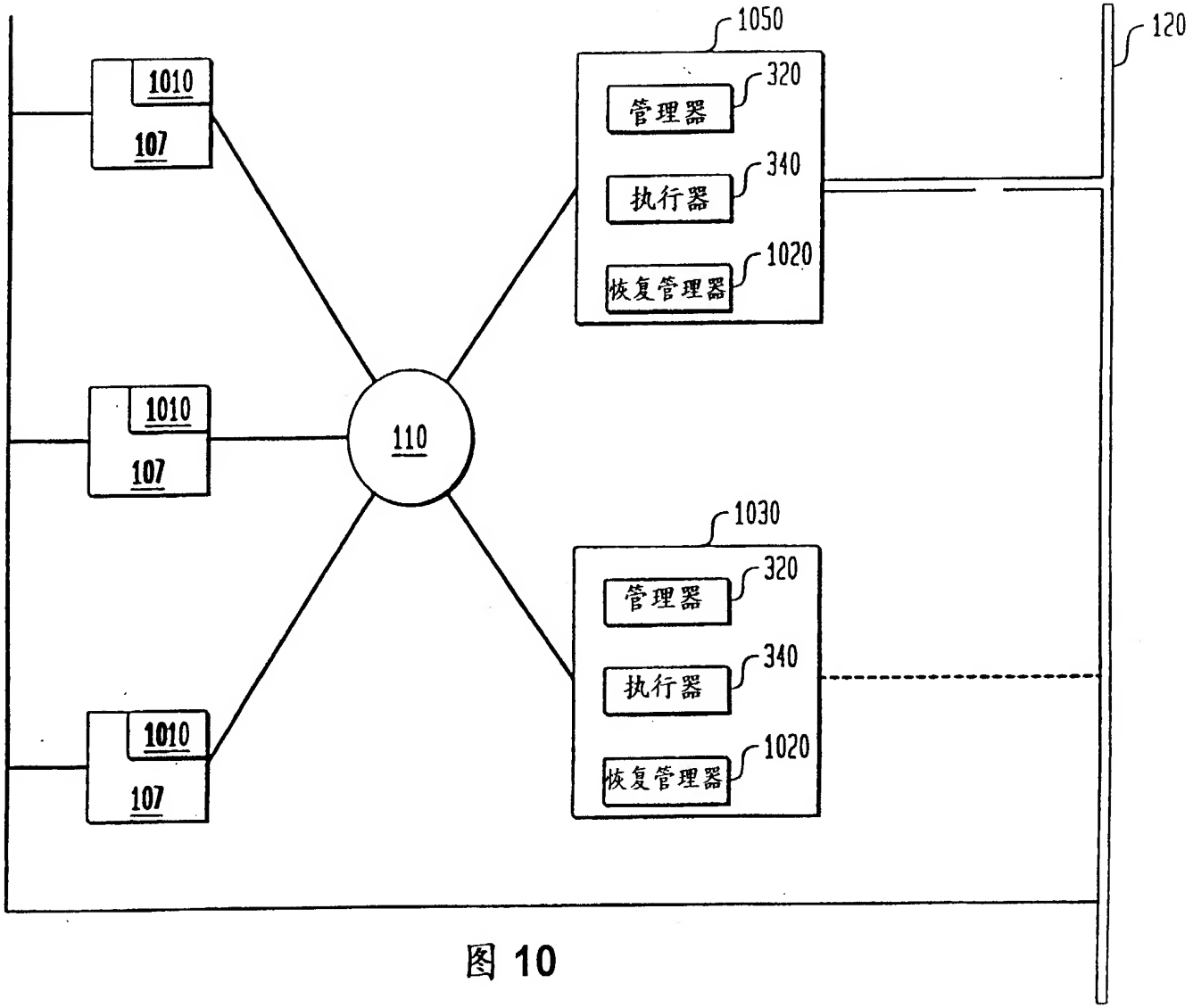

1010
107
1010
107
1010
107
110
1050
管理器
320
执行器
340
恢复管理器
1020
1030
管理器
320
执行器
340
恢复管理器
1020
120

图 10

The Delphion Integrated View: INPADOC Record

Title: **CN1175147A: VIRTUAL ENCLOSED CLUSTER CAPABLE OF RECO**

Derwent Title: Computer nodes cluster boundary incoming messages routing - by dynamically altering at least one of membership in subset and number of subsets while messages are routed across boundary [Derwent Record]

Country: CN China

Kind: A Unexamined APPLIC. open to Public inspection

Inventor: **C. R. ASANASIO**; United States of America
G. S. GERDCZMIDET; United States of America
G. D. H. HUNTER; United States of America

Assignee: **INTERNATIONAL BUSINESS MACHINE CORP.** United States of America
News, Profiles, Stocks and More about this company

Published / Filed: **1998-03-04** / 1997-07-22

Application Number: **CN19979797115437**

IPC Code: Advanced: H04L 29/08; H04L 29/06;
Core: more...
IPC-7: H04L 12/12; H04L 12/24;

ECLA Code: None

Priority Number: 1996-08-23 US19960701939

INPADOC Legal Status: None **Get Now:** Family Legal Status Report

Designated Country: AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE

Family:

PDF	Publication	Pub. Date	Filed	Title
	US20020100050A1	2002-11-07	1999-04-09	SYSTEM AND METHOD FOR PROV DYNAMICALLY ALTERABLE COMP CLUSTERS FOR MESSAGE ROUTI
	US6496866	2002-12-17	1999-04-09	System and method for providing dyr alterable computer clusters for mess
	US5918017	1999-06-29	1996-08-23	System and method for providing dyr alterable computer clusters for mess
	KR025562681	2000-05-01	1997-06-17	RECOVERABLE VIRTUAL ENCAPS CLUSTER
	JP10093655A2	1998-04-10	1997-08-20	METHOD FOR DESIGNATING WITI INCOMING MESSAGE AND SYSTE
	JP03462465B2	2003-09-29	1997-08-20	

☐	EP0838931A3	2003-12-03	1997-08-06	Recoverable virtual encapsulated clu
☐	EP0838931A2	1998-04-29	1997-08-06	Recoverable virtual encapsulated clu
☑	**CN1175147A**	1998-03-04	1997-07-22	VIRTUAL ENCLOSED CLUSTER CA RECOVERY
☑	CN1146180C	2004-04-14	1997-07-22	Virtual enclosed cluster capable of re
10 family members shown above				

Other Abstract Info: DERABS G98-232954